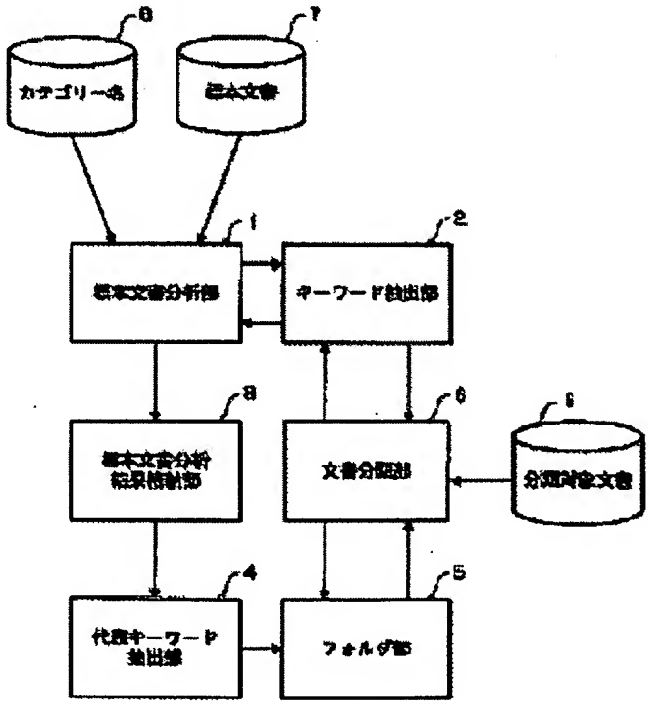
DOCUMENT SORTING DEVICE AND MACHINE-READABLE RECORDING MEDIUM RECORDING PROGRAM

Publication number: JP11025121
Publication date: 1999-01-29
Inventor: NAGATSUMA HIDEAKI
Applicant: NIPPON ELECTRIC CO
Classification:
- international: ***G06F17/30;*** **G06F17/30;** (IPC1-7): G06F17/30
- European:
Application number: JP19970189040 19970630
Priority number(s): JP19970189040 19970630

Abstract of **JP11025121**

PROBLEM TO BE SOLVED: To prevent the occurrence of events disabling preparation of a sorting dictionary and to quickly prepare the sorting dictionary in a document sorting device for preparing a sorting dictionary and for sorting a document to be sorted in each category by using the dictionary. SOLUTION: A user inputs plural sample documents 7 and a category name 8 to which these documents 7 belong in each of plural categories to which documents are to be sorted. A sample document analysis part 1 extracts a keyword included in each sample document 7 by the use of a keyword extraction part 2 and stores a pair of the extracted keyword and its corresponding category name 8 in a storing part 3 to store a sample document analysis result. A representative keyword extraction part 4 extracts a keyword included in all sample documents 7 belonging to each category as a representative keyword based on the pair of the keyword and the category name stored in the storing part 3, and stores a pair of the extracted representative keyword and the category name in a holder part 5.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－25121

(43)公開日　平成11年(1999)１月29日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/401	３１０Ｄ
			３１０Ａ

審査請求　有　　請求項の数５　ＦＤ　（全 10 頁）

(21)出願番号　特願平9－189040

(22)出願日　平成９年(1997)６月30日

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号

(72)発明者　長妻　秀明
東京都港区芝五丁目７番１号　日本電気株式会社内

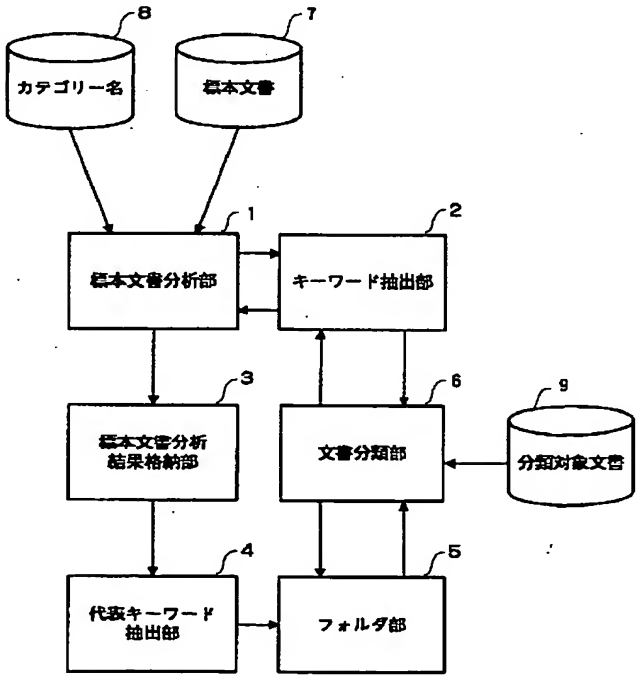
(74)代理人　弁理士　境　廣巳

(54)【発明の名称】　文書分類装置及びプログラムを記録した機械読み取り可能な記録媒体

(57)【要約】

【課題】　分類用辞書を作成し、それを用いて分類対象文書をカテゴリー別に分類する文書分類装置に於いて、分類用辞書作成不能となる事態の発生を極力抑え、且つ分類用辞書を短時間で作成できるようにする。

【解決手段】　ユーザは、分類先となる複数のカテゴリーそれぞれについて、複数の標本文書７とそれが属するカテゴリー名８を入力する。標本文書分析部１は、キーワード抽出部２を用いて標本文書７に含まれているキーワードを抽出し、抽出したキーワードとカテゴリー名８とを対にして標本文書分析結果格納部３に格納する。代表キーワード抽出部４は、標本文書分析結果格納部３中のキーワードとカテゴリー名との対に基づいて、各カテゴリー毎に、そのカテゴリーに属する全ての標本文書７に含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリー名とを対にしてフォルダ部５に格納する。



1

【特許請求の範囲】

【請求項１】　分類用辞書と、

標本文書分析結果格納部と、

文書の分類先となる複数のカテゴリーそれぞれについて、そのカテゴリーに属する複数の標本文書を入力して各標本文書に含まれるキーワードを抽出し、更に、各標本文書毎に抽出したキーワードとその標本文書が属するカテゴリーのカテゴリー名とを対にして前記標本文書解析結果格納部に格納する標本文書分析部と、

前記標本文書分析結果格納部に格納されている各標本文書のキーワードとカテゴリー名との対に基づいて、各カテゴリー毎に、そのカテゴリーに属する全ての標本文書に含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリー名とを対にして前記分類用辞書に格納する代表キーワード抽出部と、

前記分類用辞書を使用して分類対象文書をカテゴリー別に分類する文書分類部とを備えたことを特徴とする文書分類装置。

【請求項２】　前記代表キーワード抽出部の代わりに、

前記標本文書分析結果格納部に格納されている各標本文書のキーワードとカテゴリー名との対と、ユーザによって設定された代表キーワード抽出基準とに基づいて、各カテゴリー毎に、そのカテゴリーに属する全ての標本文書の内の、前記代表キーワード抽出基準によって示される割合以上の標本文書に含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリー名とを対にして前記分類用辞書に格納する拡張代表キーワード抽出部を備えたことを特徴とする請求項１記載の文書分類装置。

【請求項３】　前記文書分類部は、ユーザによって入力された分類対象文書に含まれるキーワードを抽出し、更に、前記分類用辞書を参照して前記抽出したキーワードと一致する代表キーワードを最も多く含むカテゴリーを前記分類対象文書の分類先とする構成を備えたことを特徴とする請求項１または２記載の文書分類装置。

【請求項４】　分類用辞書と標本文書分析結果格納部とを備えたコンピュータを、

文書の分類先となる複数のカテゴリーそれぞれについて、そのカテゴリーに属する複数の標本文書を入力して各標本文書に含まれるキーワードを抽出し、更に、各標本文書毎に抽出したキーワードとその標本文書が属するカテゴリーのカテゴリー名とを対にして前記標本文書解析結果格納部に格納する標本文書分析部、

前記標本文書分析結果格納部に格納されている各標本文書のキーワードとカテゴリー名との対に基づいて、各カテゴリー毎に、そのカテゴリーに属する全ての標本文書に含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリー名とを対にして前記分類用辞書に格納する代表キーワード抽出部、

前記分類用辞書を使用して分類対象文書をカテゴリー別

2

に分類する文書分類部として機能させるためのプログラムを記録した機械読み取り可能な記録媒体。

【請求項５】　分類用辞書と標本文書分析結果格納部とを備えたコンピュータを、

文書の分類先となる複数のカテゴリーそれぞれについて、そのカテゴリーに属する複数の標本文書を入力して各標本文書に含まれるキーワードを抽出し、更に、各標本文書毎に抽出したキーワードとその標本文書が属するカテゴリーのカテゴリー名とを対にして前記標本文書解析結果格納部に格納する標本文書分析部、

前記標本文書分析結果格納部に格納されている各標本文書のキーワードとカテゴリー名との対と、ユーザによって設定された代表キーワード抽出基準とに基づいて、各カテゴリー毎に、そのカテゴリーに属する全ての標本文書の内の、前記代表キーワード抽出基準によって示される割合以上の標本文書に含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリー名とを対にして前記分類用辞書に格納する拡張代表キーワード抽出部、

前記分類用辞書を使用して分類対象文書をカテゴリー別に分類する文書分類部として機能させるためのプログラムを記録した機械読み取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、電子化された文書をカテゴリー別に自動的に分類する技術に関する。

【０００２】

【従来の技術】ワードプロセッサ等によって作成された電子化された文書をカテゴリー別に分類して保存するために、文書をカテゴリー別に分類する際に使用する分類用辞書を、分類先のカテゴリーが決まっている複数の標本文書を用いて作成し、その後、作成した分類用辞書を用いて分類対象文書をカテゴリー別に分類するということが従来から行われている。

【０００３】この種の従来の技術としては、例えば、特開平６－３４８７５５号公報に記載されている技術がある。ここでは、各カテゴリーの標本文書内の単語を検出し、唯一のカテゴリーのみに出現した単語をそのカテゴリーのキーワードとして分類用辞書に登録することにより、分類用辞書を作成するようにしている。また、この技術では、分類対象文書中の単語を検出し、更に、分類用辞書に登録済みのキーワードとの一致数を検出し、一致数が最も多かったカテゴリーを分類先とするようにしている。

【０００４】また、これ以外にも、例えば、特開平１－１８８９３４号公報に記載された技術も従来から知られている。この技術では、複数の標本文書からキーワードを抽出して各キーワードの出現頻度を調べ、更に、カイ二乗計算等の複雑な計算を行って各キーワードの各カテゴリーへの貢献度を示す得点を算出することにより、得

点表（分類用辞書）を作成するようにしている。また、この技術では、分類対象文書からキーワードを抽出し、次いで、抽出した各キーワードに対応する得点を得点表から入力して分類対象文書の各カテゴリー毎の得点を計算し、最も得点の多かったカテゴリーを分類対象文書の分類先とするようにしている。

【０００５】

【発明が解決しようとする課題】上述した従来の技術の内、特開平６－３４８７５５号公報に記載されている技術では、各カテゴリーの標本文書内の単語を検出し、唯一のカテゴリーのみに出現した単語をそのカテゴリーのキーワードとして分類用辞書に登録することにより、分類用辞書を作成しているため、分類用辞書を作成できない場合があるという問題があった。つまり、特定のカテゴリーのみに現れるような単語が存在しない場合は、そのカテゴリーについてのキーワードが存在しないことになってしまうため、分類用辞書を作成することができなくなってしまう。

【０００６】また、特開平１－１８８９３４号に記載されている技術では、各キーワードの出現頻度に基づいてカイ二乗計算等の複雑な計算を行うことにより分類用辞書を作成しているため、分類用辞書の作成に多くの時間がかかってしまうという問題がある。尚、出現頻度の高いキーワードが必ずしも重要度が高いキーワードであるわけではないので、上述したような複雑な計算を行って分類用辞書を作成するようにしてもあまり意味がない。

【０００７】そこで、本発明の目的は、分類用辞書作成不能となる事態の発生を極力抑え、、且つ分類用辞書を短時間で作成できるようにすることにある。

【０００８】

【課題を解決するための手段】本発明の文書分類装置は、上記目的を達成するため、分類用辞書と、標本文書分析結果格納部と、文書の分類先となる複数のカテゴリーそれぞれについて、そのカテゴリーに属する複数の標本文書を入力して各標本文書に含まれるキーワードを抽出し、更に、各標本文書毎に抽出したキーワードとその標本文書が属するカテゴリーのカテゴリー名とを対にして前記標本文書解析結果格納部に格納する標本文書分析部と、前記標本文書分析結果格納部に格納されている各標本文書のキーワードとカテゴリー名との対に基づいて、各カテゴリー毎に、そのカテゴリーに属する全ての標本文書に含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリー名とを対にして前記分類用辞書に格納する代表キーワード抽出部と、前記分類用辞書を使用して分類対象文書をカテゴリー別に分類する文書分類部とを備えている。

【０００９】この構成に於いては、標本文書分析部が、文書の分類先となる複数のカテゴリーそれぞれについて、そのカテゴリーに属する複数の標本文書を入力して各標本文書に含まれるキーワードを抽出し、更に、各標本文書毎に抽出したキーワードとその標本文書が属するカテゴリーのカテゴリー名とを対にして標本文書解析結果格納部に格納し、代表キーワード抽出部が、標本文書分析結果格納部に格納されている各標本文書のキーワードとカテゴリー名との対に基づいて、各カテゴリー毎に、そのカテゴリーに属する全ての標本文書に含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリー名とを対にして分類用辞書に格納し、文書分類部が分類用辞書を使用して分類対象文書をカテゴリー別に分類する。

【００１０】また、本発明の文書分類装置は、分類用辞書を作成できなくなるという事態の発生を更に少なくできるようにするため、代表キーワード抽出部の代わりに、標本文書分析結果格納部に格納されている各標本文書のキーワードとカテゴリー名との対と、ユーザによって設定された代表キーワード抽出基準とに基づいて、各カテゴリー毎に、そのカテゴリーに属する全ての標本文書の内の、前記代表キーワード抽出基準によって示される割合以上の標本文書に含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリー名とを対にして前記分類用辞書に格納する拡張代表キーワード抽出部を備えている。

【００１１】この構成に於いては、拡張代表キーワード抽出部が、標本文書分析結果格納部に格納されている各標本文書のキーワードとカテゴリー名との対と、ユーザによって設定された代表キーワード抽出基準とに基づいて、各カテゴリー毎に、そのカテゴリーに属する全ての標本文書の内の、代表キーワード抽出基準によって示される割合以上の標本文書に含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリー名とを対にして分類用辞書に格納する。

【００１２】

【発明の実施の形態】次に本発明の実施の形態について図面を参照して詳細に説明する。

【００１３】図１は本発明にかかる文書分類装置の一実施例のブロック図であり、ユーザによって標本文書７とそれが属するカテゴリーのカテゴリー名８とが入力される標本文書分析部１と、キーワード抽出部２と、標本文書分析結果格納部３と、代表キーワード抽出部４と、フォルダ部５と、ユーザによって分類対象文書９が入力される文書分類部６とを備えている。

【００１４】フォルダ部５は、図２に示すように、カテゴリー名格納部５１と、文書格納部５２と、代表キーワード格納部５３とから構成されている。そして、カテゴリー名格納部５１と代表キーワード格納部５３とにより分類用辞書が構成される。

【００１５】カテゴリー名格納部５１には、文書の分類先とする全てのカテゴリーのカテゴリー名が格納される。文書格納部５２には、カテゴリー名格納部５１に格納されているカテゴリー名が示すカテゴリーに属する文

5

書が格納される。代表キーワード格納部5-3には、カテゴリー名格納部５１に格納されているカテゴリー名が示すカテゴリーを代表する代表キーワードが格納される。尚、初期状態に於いては、フォルダ部５の各部５１～５３は、情報が全く格納されていない状態になっている。

【００１６】図２の例は、文書の分類先とするカテゴリーが「入会手続」，「顧客管理」，「操作方法」，「カタログ請求」，「支払請求」であり、各カテゴリーを代表する代表キーワードがそれぞれ「入会，手続，住所，名前」，「住所，電話，変更」，「変更，方法」，「カタログ，住所，送付」，「料金，支払」であることを示している。更に、図２の例は、各カテゴリーに属する文書としてそれぞれ「文書１，文書６，文書９」，「文書３，文書７，文書１０」，「文書２，文書４，文書１２」，「文書８，文書１３」，「文書５，文書１１，文書１４」が格納されていることを示している。

【００１７】標本文書分析部１は、ユーザによって標本文書７とカテゴリー名８が入力されると、キーワード抽出部２に標本文書７を渡してキーワードを抽出させる機能，キーワード抽出部２によって抽出されたキーワードとカテゴリー名８とを対にして標本文書分析結果格納部３に格納する機能を有する。

【００１８】キーワード抽出部２は、標本文書分析部１から標本文書７が渡された場合は、標本文書７に含まれているキーワードを抽出して標本文書分析部１に返し、文書分類部６から分類対象文書９が渡された場合は、分類対象文書９に含まれているキーワードを抽出して文書分類部６に返す機能を有する。尚、キーワードの抽出方法は、キーワードを抽出できればどのような方法でも良く、例えば、キーワード抽出対象文書に対して形態素解析を行い、それに含まれている名詞をキーワードとして抽出するといった方法や、その他の周知の方法をとることができる。

【００１９】代表キーワード抽出部４は、標本文書分析結果格納部３に格納されている各標本文書のキーワードとカテゴリー名との対に基づいて、各カテゴリー毎に、そのカテゴリーに属する全ての標本文書に含まれているキーワードを代表キーワードとして抽出する機能，抽出した各カテゴリーの代表キーワードとそれが代表するカテゴリーのカテゴリー名とを対応付けてフォルダ部５の代表キーワード格納部５３，カテゴリー名格納部５１に格納する機能を有する。

【００２０】文書分類部６は、ユーザによって分類対象文書９が入力されると、それをキーワード抽出部２に渡して分類対象文書９に含まれるキーワードを抽出させる機能や、フォルダ部５のカテゴリー名格納部５１，代表キーワード格納部５３（上述したように、両者によって分類用辞書が構成される）を参照して上記抽出したキーワードと一致する代表キーワードを最も多く含むカテゴリーを分類対象文書９の分類先のカテゴリーとする機能

6

や、分類対象文書９を分類先のカテゴリーのカテゴリー名に対応付けてフォルダ部５に格納する機能を有する。

【００２１】図３は標本文書分析部１の処理例を示す流れ図、図４は代表キーワード抽出部４の処理例を示す流れ図、図５は文書分類部６の処理例を示す流れ図であり、以下各図を参照して本実施例の動作を説明する。

【００２２】ユーザは、先ず、分類用辞書を作成するために、或るカテゴリーに属する標本文書７と、それが属するカテゴリーのカテゴリー名８とを標本文書分析部１に入力する。尚、標本文書７は、既に作成済みの文書の中から適切な文書を選び出すようにした方が、作業を効率的に行うことができるが、新たに作成するようにしても構わない。

【００２３】標本文書分析部１は、標本文書７，カテゴリー名８が入力されると、図３の流れ図に示すように、標本文書７をキーワード抽出部２に渡す（Ｓ３１）。これにより、キーワード抽出部２は、標本文書７に含まれている全てのキーワードを抽出して標本文書分析部１に返す。

【００２４】キーワード抽出部２からキーワードを受け取ると（Ｓ３２）、標本文書分析部１は、ユーザによって入力されたカテゴリー名８とキーワード抽出部２から渡されたキーワードとを対にしたレコードを標本文書分析結果格納部３に格納する（Ｓ３３）。

【００２５】今、例えば、ユーザが、カテゴリー名８が「入会手続」の標本文書７を入力したとすると、標本文書分析部１は、先ず、標本文書７をキーワード抽出部２に渡してそれに含まれる全てのキーワードを抽出させる（Ｓ３１）。キーワード抽出部２で抽出されたキーワードが「入会，手続，住所，料金，電話」であったとすると、標本文書分析部１は、カテゴリー名「入会手続」とキーワード「入会，手続，住所，料金，電話」とを対にしたレコードを標本文書分析結果格納部３に格納する。

【００２６】ユーザは、文書の分類先とする全てのカテゴリーについて、それぞれ複数の異なる標本文書７を入力すると共にカテゴリー名８を入力する。標本文書分析部１は、ユーザによって標本文書７，カテゴリー名８が入力される毎に上述したと同様の処理を行う。

【００２７】図６は、ユーザが全ての標本文書７を入力し終わった時の標本文書分析結果格納部３の内容例を示した図である。この例は、カテゴリー名「入会手続」，「顧客管理」，「操作方法」，…の標本文書がそれぞれ５文書，３文書，３文書，…ずつ入力された時の例である。

【００２８】ユーザは、全ての標本文書７を入力すると、代表キーワード抽出部４を動作させる。

【００２９】これにより、代表キーワード抽出部４は、図４に示すように、標本文書分析結果格納部３に格納されているレコード中にカテゴリー名の１つに注目する（Ｓ４１）。その後、代表キーワード抽出部４は、注目

したカテゴリー名を含む全てのレコードを標本文書分析結果格納部３から取り出す（Ｓ４２）。

【００３０】今、例えば、標本文書分析結果格納部３の内容が図６に示すもので、注目しているカテゴリー名が「入会手続」であったとすると、代表キーワード抽出部４は、図７に示す５個のレコードを標本文書分析結果格納部３から取り出すことになる。

【００３１】その後、代表キーワード抽出部４は、Ｓ４２で取り出した全てのレコードに含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリー名とを対応付けてフォルダ部５の代表キーワード格納部５３，カテゴリー名格納部５１に格納する（Ｓ４３，Ｓ４４）。

【００３２】今、例えば、Ｓ４２で図７に示した５個のレコードを標本文書分析結果格納部３から取り出したとすると、代表キーワード抽出部４は、５個のレコードの全てに含まれている３個のキーワード「入会，手続，住所」を代表キーワードとして抽出し、抽出した代表キーワード「入会，手続，住所」とカテゴリー名「入会手続」とをフォルダ部５の代表キーワード格納部５３，カテゴリー名格納部５１に格納することになる（Ｓ４３，Ｓ４４）。

【００３３】代表キーワード抽出部４は、未注目のカテゴリー名がなくなるまで（Ｓ４５がＮＯとなるまで）、上記した処理を繰り返し行う。図８は、未注目のカテゴリー名がなくなるまで、上記した処理を行った時にフォルダ部５の内容例を示した図であり、この例は、カテゴリー名「入会手続」，「顧客管理」，「操作方法」，「カタログ請求」，「支払請求」のカテゴリーの代表キーワードがそれぞれ「入会，手続，住所」，「住所，電話，変更」，「変更，方法」，「カタログ，住所，送付」，「料金，支払」であることを示している。

【００３４】フォルダ部５のカテゴリー名格納部５１，代表キーワード格納部５３に、上述したようにしてカテゴリー名，代表キーワードを格納し、分類を行うための準備が完了すると、ユーザは、分類対象文書９を文書分類部６に入力する。

【００３５】文書分類部６は、分類対象文書９が入力されると、図５の流れ図に示すように、分類対象文書９をキーワード抽出部２に渡す（Ｓ５１）。これにより、キーワード抽出部２は、分類対象文書９に含まれているキーワードを全て抽出し、抽出したキーワードを全て文書分類部６に返す。

【００３６】文書分類部６は、キーワード抽出部２からキーワードを受け取ると（Ｓ５２）、フォルダ部５のカテゴリー名格納部５１，代表キーワード格納部５３を参照し、キーワードと一致する代表キーワードを最も多く含むカテゴリー名を求め（Ｓ５３）、その後、Ｓ５３で求めたカテゴリー名と対応付けてフォルダ部５の文書格納部５２に分類対象文書９を格納する（Ｓ５４）。

【００３７】今、例えば、分類対象文書９として、「入会したいので、手続に必要な資料を下記の住所まで送付して下さい。」が入力され、フォルダ部５の内容が図８に示すものであったとすると、、以下の処理が行われることになる。

【００３８】先ず、文書分類部６が、入力された分類対象文書「入会したいので、手続に必要な資料を下記の住所まで送付して下さい。」をキーワード抽出部２に渡す（Ｓ５１）。

【００３９】これにより、キーワード抽出部２がキーワードを抽出し、抽出したキーワードを文書分類部６に返す。今、例えば、キーワードとして「入会，手続，資料，住所，送付」が抽出されたとする。

【００４０】文書分類部６は、キーワード抽出部２から上記したキーワード「入会，手続，資料，住所，送付」を受け取ると（Ｓ５２）、キーワードと一致する代表キーワードを最も多く含むカテゴリー名を求める（Ｓ５３）。図８の例では、上記したキーワードと一致する代表キーワードの数は、「入会手続」で３個，「顧客管理」で１個，「操作方法」で０個，「カタログ請求」で２個，「支払請求」で０個となるので、最も一致数の多い「入会手続」が選ばれることになる。

【００４１】その後、文書分類部６は、フォルダ部５中の文書格納部５２の、カテゴリー名「入会手続」と対応する部分に、入力された分類対象文書「入会したいので、手続に必要な資料を下記の住所まで送付して下さい。」を格納する。

【００４２】図９は、本発明にかかる文書分類装置の他の実施例のブロック図であり、図１に示した文書分類装置との相違点は、代表キーワード抽出部４の代わりに拡張代表キーワード抽出部４’を備えた点、代表キーワード抽出基準１０を入力する代表キーワード抽出基準入力部１１を追加した点である。尚、図９に於いて、他の図１と同一符号は、同一部分を表している。

【００４３】代表キーワード抽出基準入力部１１は、ユーザが設定した代表キーワード抽出基準１０を入力する機能を有する。ここで、代表キーワード抽出基準１０は、代表キーワードの抽出基準を示すものであり、Ａ／Ｂ（Ａ≦Ｂ）の形式を有する。つまり、或るカテゴリーに属する複数の標本文書の内の、代表キーワード抽出基準１０によって示される割合Ａ／Ｂ以上の標本文書に含まれているキーワードを代表キーワードにすることを指示するものである。尚、代表キーワード抽出基準１０の値は、変更可能なものである。

【００４４】拡張代表キーワード抽出部４’は、標本文書分析結果格納部３に格納されている図６に示すような各レコードと、代表キーワード抽出基準入力部１１によって入力された代表キーワード抽出基準１０とに基づいて、各カテゴリー毎に、そのカテゴリーに属する全ての文書の内の、代表キーワード抽出基準１０によって示さ

れる割合を超える標本文書に含まれるキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリー名とを対応付けてフォルダ部５に格納する機能を有する。

【００４５】図１０は拡張代表キーワード抽出部４’の処理例を示す流れ図であり、以下各図を参照して本実施例の動作を説明する。

【００４６】ユーザは、先ず、分類用辞書を作成するために、或るカテゴリーに属する標本文書７と、それが属するカテゴリーのカテゴリー名８とを標本文書分析部１に入力する。

【００４７】標本文書分析部１は、標本文書７，カテゴリー名８が入力される毎に前述したと同様の処理を行い（図３，Ｓ３１～Ｓ３３）、標本文書分析結果格納部３に図６に示すようなレコードを格納する。

【００４８】ユーザは、分類用辞書を作成するために用意した全ての標本文書７を入力すると、代表キーワード抽出基準１０を設定する。これにより、代表キーワード抽出基準入力部１１が、代表キーワード抽出基準１０を拡張代表キーワード抽出部４’に入力する。

【００４９】代表キーワード抽出基準１０が入力されると、拡張代表キーワード抽出部４’は、図１０の流れ図に示すように、標本文書分析結果格納部３に格納されているレコードに含まれているカテゴリー名の内の１つに注目する（Ｓ１０１）。

【００５０】その後、拡張代表キーワード抽出部４’は、注目したカテゴリー名を含む全てのレコードを標本文書分析結果格納部３から取り出す（Ｓ１０２）。

【００５１】次いで、拡張代表キーワード抽出部４’は、Ｓ１０２で取り出したレコードの内の、代表キーワード抽出基準１０によって示される割合以上のレコードに含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリー名とを対応付けてフォルダ部５の代表キーワード格納部５３，カテゴリー名格納部５１に格納する（Ｓ１０３，Ｓ１０４）。その後、拡張代表キーワード抽出部４’は、未注目のレコードがなくなるまで（Ｓ１０５がＮＯとなるまで）、上述した処理を繰り返し行う。

【００５２】今、例えば、標本文書分析結果格納部３の内容が図６に示すものであり、ユーザが設定した代表キーワード抽出基準１０が「３／５」であったとすると、以下の処理が行われる。

【００５３】先ず、標本文書分析結果格納部３に格納されているカテゴリー名の内の１つ（「入会手続」とする）に注目する（Ｓ１０１）。次に、標本文書分析結果格納部３からカテゴリー名が「入会手続」となっているレコードを全て取り出す（Ｓ１０２）。この例の場合、図７に示す５個のレコードが取り出される。

【００５４】その後、図７に示した５個のレコードの内の、代表キーワード抽出基準１０によって示される割合「３／５」以上のレコードに含まれるキーワードを代表キーワードとして抽出する（Ｓ１０３）。この例の場合、キーワードが含まれる割合は、「入会」＝５／５，「手続」＝５／５，「住所」＝５／５，「料金」＝４／５，「名前」＝４／５，「電話」＝２／５，「ＦＡＸ」＝１／５，「方法」＝１／５であるので、拡張代表キーワード抽出部４’は、「入会」，「手続」，「住所」，「料金」，「名前」の５つを代表キーワードとして抽出する。

【００５５】その後、拡張代表キーワード抽出部４’は、カテゴリー名「入会手続」と代表キーワード「入会，手続，住所，料金，名前」とを対応付けてフォルダ部５に格納する。

【００５６】以下、拡張代表キーワード抽出部４’は、未注目のレコードがなくなるまで（Ｓ１０５がＮＯとなるまで）、上述した処理を繰り返し行う。図１１は、未注目のレコードがなくなるまで、上述した処理を行った時のフォルダ部５の内容を示した図である。

【００５７】フォルダ部５のカテゴリー名格納部５１，代表キーワード格納部５３に、上述したようにしてカテゴリー名，代表キーワードを格納し、分類を行うための準備が完了すると、ユーザは、分類対象文書９を文書分類部６に入力する。

【００５８】文書分類部６は、分類対象文書９が入力されると、前述したと同様にして、分類対象文書９を文書格納部５２に該当する部分に格納する。

【００５９】図１２は、図１及び図９に示した文書分類装置のハードウェア構成を示したブロック図であり、コンピュータ１２１と、記録媒体１２２と、記憶装置１２３とを備えている。記録媒体１２２は、磁気ディスク，半導体メモリ，その他の記録媒体である。

【００６０】図１に示した文書分類装置を実現する場合には、記録媒体１２２に記録された文書分類用プログラムがコンピュータ１２１によって読み込まれ、コンピュータ１２１の動作を制御することで、コンピュータ１２１上に図１に示した標本文書分析部１，キーワード抽出部２，代表キーワード抽出部４，文書分類部６を実現する。尚、標本文書分析結果格納部３，フォルダ部５は、記憶装置１２３上に構成される。

【００６１】また、図９に示した文書分類装置を実現する場合には、記録媒体１２２に記録された文書分類用プログラムがコンピュータ１２１によって読み込まれ、コンピュータ１２１の動作を制御することで、コンピュータ１２１上に図９に示した標本文書分析部１，キーワード抽出部２，拡張代表キーワード抽出部４’，文書分類部６，代表キーワード抽出基準入力部１１を実現する。尚、標本文書分析結果格納部３，フォルダ部５は、記憶装置１２３上に構成される。

【００６２】

【発明の効果】以上説明したように、本発明の文書分類

装置は、標本文書分析結果格納部に格納されている各標本文書のキーワードとカテゴリー名との対に基づいて、各カテゴリー毎に、そのカテゴリーに属する全ての標本文書に含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリー名とを対にして分類用辞書に格納することにより分類用辞書を作成する代表キーワード抽出部を備えているので、唯一のカテゴリーのみに出現した単語をそのカテゴリーのキーワードとして分類用辞書に登録することにより分類用辞書を作成するようにしていた従来技術に比較して、分類用辞書が作成不能となる危険性を少なくすることができ、且つカイ二乗計算等の複雑な計算を行うことにより分類用辞書を作成していた従来に技術に比較して分類用辞書の作成時間を短くすることができる。

【００６３】また、本発明の文書分類装置は、ユーザによって指定された代表キーワード抽出基準と、標本文書解析結果格納部の格納されている各標本文書のキーワードとカテゴリー名との対に基づいて、各カテゴリー毎に、そのカテゴリーに属する全ての標準文書の内の、代表キーワード抽出基準によって示される割合以上の標本文書に含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリー名とを対にして分類用辞書に格納することにより分類用辞書を作成するようにしているので、分類用辞書が作成不能となる危険性を更に少なくすることができる。更に、或るカテゴリーを示す重要なキーワードが全ての標本文書に含まれていなかったとしても、そのキーワードを代表キーワードとして含む分類用辞書を作成することができる。

【図面の簡単な説明】

【図１】本発明の一実施例のブロック図である。

【図２】フォルダ部５の内容例を示す図である。

【図３】標本文書分析部１の処理例を示す流れ図である。

【図４】代表キーワード抽出部４の処理例を示す流れ図である。

【図５】文書分類部６の処理例を示す流れ図である。

【図６】標本文書分析結果格納部３の内容例を示す図である。

【図７】代表キーワード抽出部４が標本文書分析結果格納部３から取り出した、カテゴリー名が同一のレコードを示した図である。

【図８】カテゴリー名格納部５１，代表キーワード格納部５３にカテゴリー名，代表キーワードを格納し終わった時のフォルダ部５の内容例を示す図である。

【図９】本発明の他の実施例のブロック図である。

【図１０】拡張代表キーワード抽出部４’の処理例を示す流れ図である。

【図１１】カテゴリー名格納部５１，代表キーワード格納部５３にカテゴリー名，代表キーワードを格納し終わった時のフォルダ部５の内容例を示す図である。

【図１２】文書分類装置のハードウェア構成の一例を示すブロック図である。

【符号の説明】

１…標本文書分析部
２…キーワード抽出部
３…標本文書分析結果格納部
４…代表キーワード抽出部
４’…拡張代表キーワード抽出部
５…フォルダ部
５１…カテゴリー名格納部
５２…文書格納部
５３…代表キーワード格納部
６…文書分類部
７…標本文書
８…カテゴリー名
９…分類対象文書
１０…代表キーワード抽出基準
１１…代表キーワード抽出基準入力部
１２１…コンピュータ
１２２…記録媒体
１２３…記憶装置

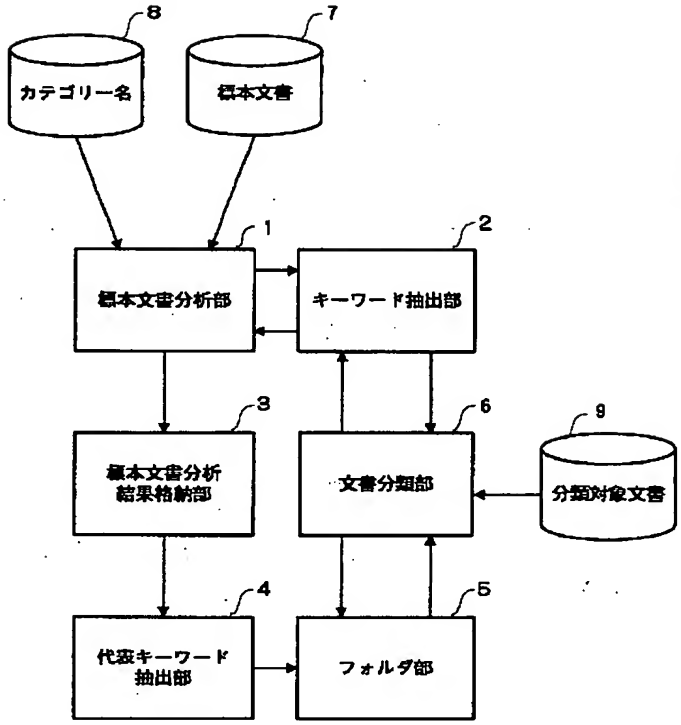
【図７】

カテゴリー名	キーワード
入会手続	入会，手続，住所，料金，電話
入会手続	入会，手続，住所，名前，料金，電話
入会手続	入会，手続，住所，名前，料金，FAX
入会手続	入会，手続，住所，名前，料金
入会手続	入会，手続，住所，名前，方法

【図８】

51 カテゴリー名	52 カテゴリーに属する文書	53 代表キーワード
入会手続		入会，手続，住所
顧客管理		住所，電話，変更
操作方法		変更，方法
カタログ請求		カタログ，住所，送付
支払請求		料金，支払

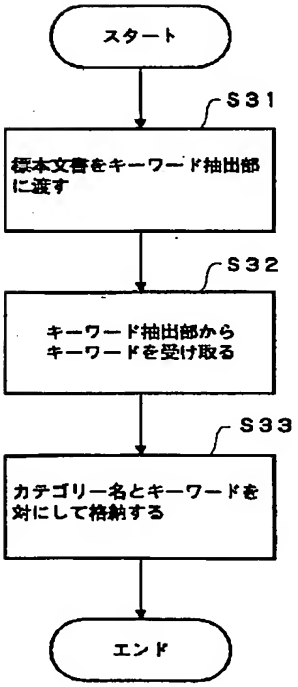
【図１】



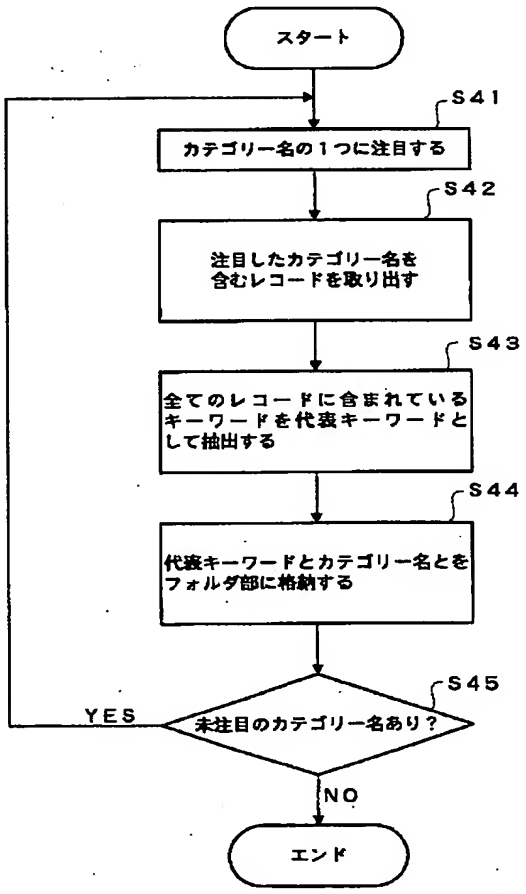
【図２】

51	52	53
カテゴリー名	カテゴリーに属する文書	代表キーワード
入会手続	文書１，文書６，文書９	入会，手続，住所，名前
顧客管理	文書３，文書７，文書10	住所，電話，変更
操作方法	文書２，文書４，文書12	変更，方法
カタログ請求	文書８，文書13	カタログ，住所，送付
支払請求	文書５，文書11，文書14	料金，支払

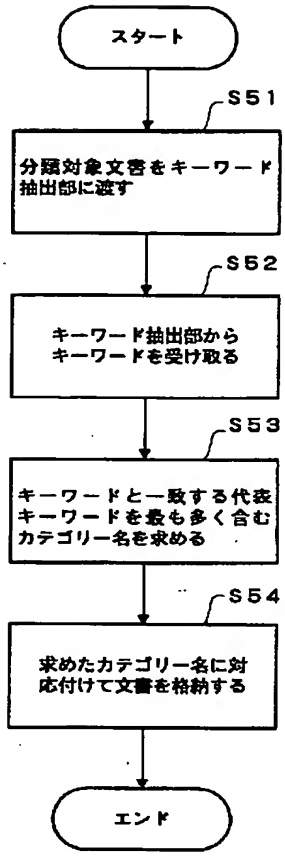
【図３】



【図４】



【図５】

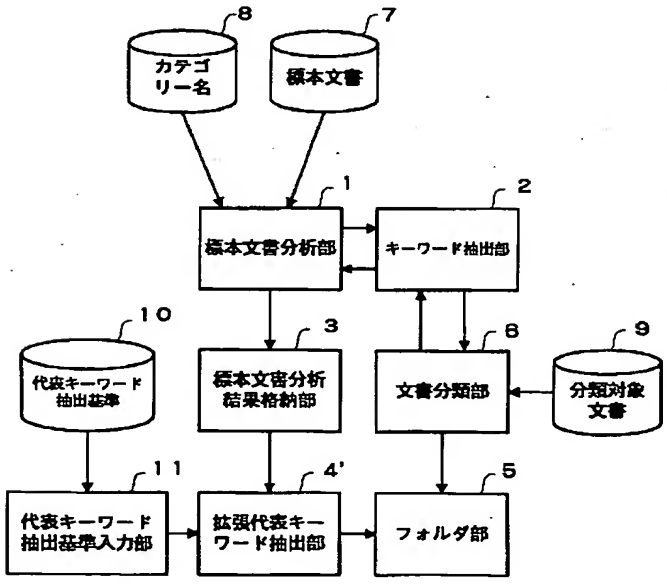


【図６】

3

カテゴリー名	キーワード
入会手続	入会，手続，住所，料金，電話
入会手続	入会，手続，住所，名前，料金，電話
入会手続	入会，手続，住所，名前，料金，FAX
入会手続	入会，手続，住所，名前，料金
入会手続	入会，手続，住所，名前，方法
顧客管理	住所，電話，変更，名前，方法
顧客管理	住所，電話，変更，名前
顧客管理	住所，電話，変更，パスワード
操作方法	変更，方法，修正，入力，出力
操作方法	変更，方法，修正，入力
操作方法	変更，方法，パスワード
⋮	

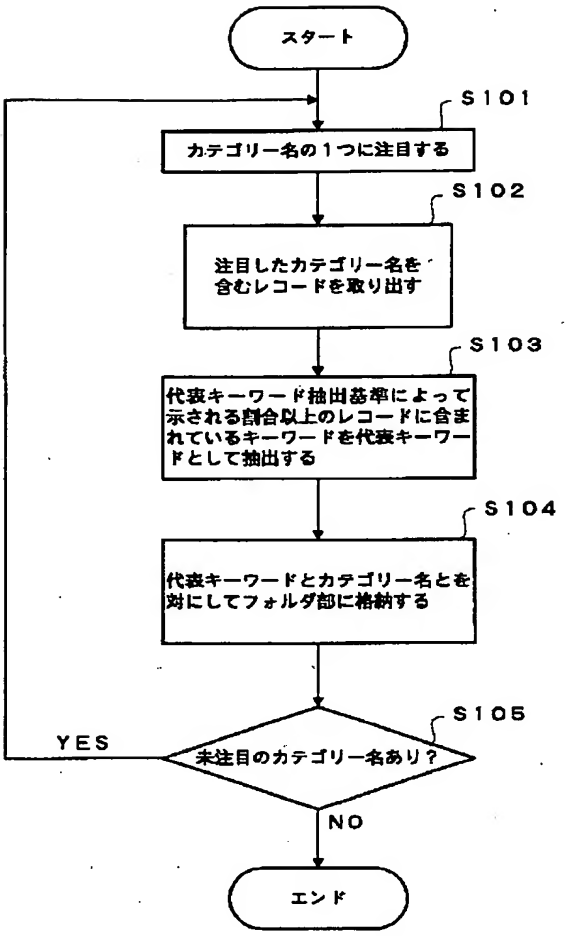
【図９】



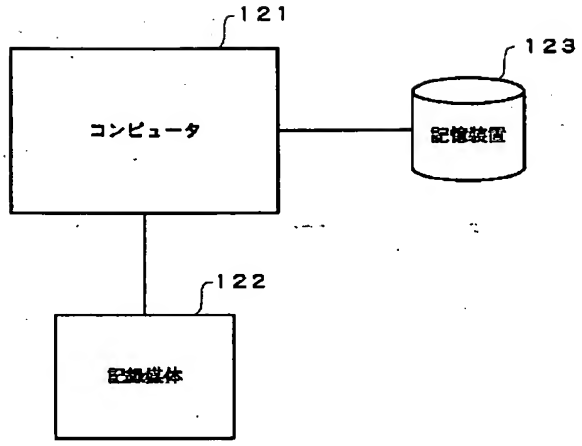
【図１１】

51 カテゴリー名	52 カテゴリーに属する文書	53 代表キーワード
入会手続		入会，手続，住所，料金，名前
顧客管理		住所，電話，変更
操作方法		変更，方法
カタログ請求		カタログ，住所，送付
支払請求		料金，支払

【図１０】


スタート
S101
カテゴリー名の１つに注目する
S102
注目したカテゴリー名を含むレコードを取り出す
S103
代表キーワード抽出基準によって示される割合以上のレコードに含まれているキーワードを代表キーワードとして抽出する
S104
代表キーワードとカテゴリー名とを対にしてフォルダ部に格納する
S105
未注目のカテゴリー名あり？
YES
NO
エンド

【図１２】


121
123
コンピュータ
記憶装置
122
記録媒体